Semantic Incompleteness in Privacy Policy Goals

Jaspreet Bhatia and Travis D. Breaux
Institute for Software Research, Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
{jbhatia, breaux}@cs.cmu.edu

***Abstract* — Companies that collect personal information online often maintain privacy policies that are required to accurately reflect their data practices and privacy goals. To be comprehensive and flexible for future practices, policies contain ambiguity that summarize practices over multiple types of products and business contexts. Ambiguity in data practice descriptions undermines policies as an effective way to communicate system design choices to users, and as a reliable regulatory mechanism. In this paper, we report an investigation to identify incompleteness by representing data practice descriptions as semantic frames. The approach is a grounded analysis to discover which data actions and semantic roles correspond are needed to construct complete data practice descriptions. Our results include 281 data action instances obtained from 202 manually annotated statements across five privacy policies. Therein, we identified 878 instances of 17 types of semantic roles. Incomplete data practice descriptions undermine user comprehension, and can affect the user's perceived privacy risk, which we measure using factorial vignette surveys. We observed that user perception of risk decreases when two roles are present in a statement: the condition under which a data action is performed, and the purpose for which the user's information is used.**

***Index Terms* — semantic frames, semantic roles, privacy risk, natural language processing, privacy.**

I. Introduction

Companies describe their data practices in privacy policies to inform users about how their data must be collected, used and transferred for the purposes embodied by the website or software. U.S. regulators may check these policies for compliance with actual data practices, when a data breach or data misuse arises. Consequently, the statements in policies represent legal requirements for software systems. Ideally, users can also use these policies to better understand what the website does with their personal information and to make informed decisions about using the services provided by the website.

A company's data practice descriptions in a privacy policy can govern multiple types of products, and both physical and virtual stores. In addition, policies are drafted to account for current practices, as well as to afford flexibility for future practices that the company envisions. In doing so, companies resort to using ambiguity in the data practice descriptions of their policies. In the worst case, this ambiguity can lead to inaccurate interpretations by users and regulators.

Privacy policy statements correspond to privacy goals and requirements. Incompleteness in requirements can lead to misunderstanding among stakeholders, wherein stakeholders have different interpretations regarding the incomplete information [11]. Incomplete privacy goals convey to developers a potentially inaccurate description of requirements that should be met by the system. Incomplete requirements are one of the most critical challenges faced by software companies and are also a frequent cause of project failures [13].

Incompleteness*,* which is a form of ambiguity, occurs in data practice descriptions when one or more policy statements do not answer all the questions that users or regulators may have regarding the company's data practices. For example, with respect to the data action "share," one could ask: what type of data is shared? With whom will the data be shared? From whom was the data collected? For what purpose is the data shared? Finally, under what conditions will the data be shared? If the data practice description does not answer one or more of these questions, the description can be considered incomplete with respect to the missing information.

Incompleteness in privacy goals and requirements can prevent users from making accurate predictions about how their data is collected, retained, shared or used by the company, consequently causing users to misestimate their personal privacy risk. For example, in the summary privacy statement "we may share your location information," the purpose for which the user's location information is shared is missing, which requires the user to make assumptions about the missing purpose. The user may assume that the sharing is undertaken for a primary purpose for which the data was collected, for example to provide services requested by the user, which leads to underestimating the risk. Alternatively, the user may assume that the shared data is used for an unstated, secondary purpose, either by a first party or third party [5]. Secondary use can lead to overestimation of the privacy risk by users, despite that the third party's data practice remains unknown.

The overestimation of privacy risk is not a favorable situation for a company, because it can lead to either the user not using a service due to fear of data misuse, or it can lead to the regulator concluding that the data practice is not in compliance with a regulation. In 2015, the social networking website and application Snapchat changed its data practice descriptions in their privacy policy concerning collection, use and retention of their user data, stating that "…we may access, review, screen, delete your content at any time and for any reason" and "…publicly display that content in any form and in any and all media or distribution methods." Such statements led users to

2332-6441/18/$31.00 ©2018 IEEE
DOI 10.1109/RE.2018.00025

worry about the ways in which their information could be collected, retained and used, since the policy was extremely permissive. This led some users to report that they had deleted their accounts[1]. In another incident, Google was warned by European regulators about vagueness in their policy concerning data retention practices and about not showing a commitment towards the European Data Protection Directive[2]. Therefore, companies should identify when a data practice is incomplete and take corrective measures to improve the description.

In this paper, we identify incompleteness by representing a data practice description (a data action) as a *semantic frame*. We construct these frames by identifying relevant questions for each data action, which we call *semantic roles* associated with the action. We propose to develop a network of semantic frames to determine the roles that are expected to complete a data practice description. In so doing, we aim to understand how roles contribute context for an action, and how policy authors choose roles when expressing privacy policies. For example, the following JCPenny privacy policy statement is annotated for semantic roles that describe the data action *collect* in Figure 1. The condition on the action collect is "when you interact with JC Penny", the object is "information," the source of the information is "you," and the purpose of collection is "to provide you services."

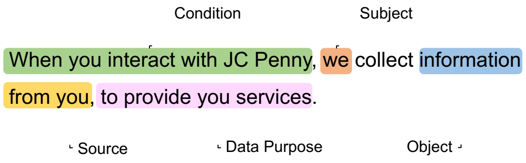

Condition
Subject
When you interact with JC Penny, we collect information
from you, to provide you services.
Source
Data Purpose
Object

Fig. 1. Example statement with annotated semantic roles

This paper is organized as follows: in Section II, we review background and related work; in Section III, we describe our approach towards building semantic frames for data practices, and our grounded analysis results; in Section IV, we present the design of user studies to measure the perceived privacy risk due to incomplete data practice descriptions and the study results; in Section V, we report the threats to validity, and in Section VI, we discuss our research questions in light of our results and future work.

II. Background and Related Work

We now review background and prior work related to semantic frames and roles in natural language, semantic frame representations for requirements, and privacy risk.

We identify incompleteness in data practices by determining which of the expected roles for a data action are missing values in data practice statements. In order to determine the expected roles that will help us better understand a data action, we need to answer questions associated with that action, such as *who performs the action* and *on what data the action was performed*, among other questions [19]. The answers to these questions can be expressed in many different ways in a statement. For example, consider the following data practice statements:

- `We collect user information.`
- `The user information is logged by us.`
- `We gather information about our users.`
- `The user provides us with their information.`

While the above statements use different action words such as *collect, log, gather,* and *provide*, and have different syntax, they have similar meaning, which is that the user information is collected by the website. One representation that permits comparison among these statements is called semantic roles [19]. Roles are considered shallow representations, because they rely only on the relationship between a given word or role value and other clauses in the statement, and not among all the words in the statement. Using semantic roles, we represent the fact that there is a *collection* action taking place, the action is being performed by the subject, *the website company*, and the object of the action is the *user information*. Semantic roles represent the relationship of the different clauses in the statement to the main action, like the subject and object [19]. The context of a data action can be expressed using different semantic roles, such as agent (who initiates and performs an action), patient (what undergoes the action and changes its state), instrument (used to carry out the action), source (where the action originated), among other roles [17].

Semantic roles that are used to describe a data action can be represented together in a knowledge representation technique known as *frames*. Minsky describes a frame as a data structure that is used to represent a stereotyped situation, such as being in a certain kind of living room [23]. Each frame is associated with *slots* or semantic roles, which are filled by *fillers* or semantic role values in specific contexts, and which help readers understand a situation in question. The values for these semantic roles can be atomic values, procedures, or pointers to other frames [23]. Frames can be used to represent knowledge in a succinct manner and to reason in an efficient way [14].

According to Fillmore's frame semantics, the meaning of a word cannot be understood in isolation, but in conjunction with the information that relates to it [18]. For example, the word "share" can be understood when we have knowledge about who is sharing, what is being shared, and with whom it is being shared. Fillmore's frame semantics are implemented in the FrameNet project [4]. The FrameNet corpus contains manually annotated, general purpose semantic frames for the English language, with semantic roles specific to a frame. The *frames* are evoked by *lexical units* which are lemmas and their part of speech. The semantic roles associated with each frame are also known as *frame elements*, which provide information about the frame. Consider the following example from the FrameNet database:

`Abby bought a car from Robin.`

In this statement, the *frame* "commerce_buy" is evoked by the *lexical unit* "bought (buy.verb)". The *frame elements* of this frame instantiated in this statement are: *buyer* (Abby), *goods* (a

[1] Sally French, "Snapchat's new 'scary' privacy policy has left users outraged," Market Watch, 2 November 2015. http://www.marketwatch.com/story/snapchats-new-scary-privacy-policy-has-left-users-outraged-2015-10-29

[2] Zack Whittaker, "Google must review privacy policy, EU data regulators rule," ZDNet, 16 October 2012. http://www.zdnet.com/article/google-must-review-privacy-policy-eu-data-regulators-rule/

car), *seller* (robin). Similar to FrameNet our frames are evoked by different categories of data actions, which represent a situation where the user's information is being acted upon by a company. We employ semantic roles that are specific to each such frame, and are instantiated when that frame is evoked. The FrameNet resource has been used for automatic semantic role labelling [12, 25]. Das et al. report an F1 score of 61.4 and 68.49 for frame identification and semantic role value identification respectively for SemEval 2007 data, and F1 score of 80.3 and 79.9 for frame identification and role value identification respectively on the FrameNet 1.5 release [12]. Semantic role labelling has been used for improving applications such as question-answering [20], recognizing textual entailment [8], information extraction [28] and in requirements engineering, to extract information from software requirements specifications [32].

In this paper, we identify the expected semantic roles for a given frame, and consequently determine when the information provided is incomplete, by identifying roles that are missing values in a given data practice statement. Our analysis in this study is limited to the contextual information provided in a single statement and we do not combine contextual information from multiple statements. Because incomplete information prevents users from having control over their information and knowing when an entity has access to their information, it can also affect a user's perception of their privacy [2]. In addition, incompleteness prevents users from knowing the potential consequences of such disclosures. Tsai et al. found that users took privacy information into consideration while making decisions about using the services of an online website and were willing to pay to protect their privacy [29]. These findings make it important to identify the privacy risk perceived by a user due to incomplete information. Furthermore, websites can provide more complete information about their data practices to help users make better decisions about using the services provided by the website.

III. Semantic Role Representation and Incompleteness

Our research questions are as follows:

RQ1. What are the different semantic roles associated with different categories of data actions?

RQ2. What are the variations in the values of the different semantic roles?

RQ3. What are the different lexical and syntactic triggers that indicate semantic role values?

RQ4. How does the presence or absence of semantic roles and their values affect the user's perception of privacy risk?

To answer the first three research questions, we manually annotated semantic roles in five privacy policies. We chose a convenience sample of five policies from the shopping domain (see Table I), wherein the companies maintain both online and "brick-and-mortar" stores.

TABLE I. Privacy Policy Dataset For Semantic Frame Study

Company Name	**Last Updated**
Barnes and Noble	08/05/2016
Costco	12/31/2013
JC Penny	09/01/2016
Lowes	08/20/2015
Overstock	06/20/2017

A. Annotating and Extracting Semantic Roles

The first three research questions concern the different semantic roles and their variations across different data actions, and the lexical and syntactic triggers that indicate semantic role values. We annotated the policies in Table I using content analysis, in which an analyst assigns codes to text from a coding frame [26]. Each coded text fragment represents an instance of the code, after which the analyst can review the coded items for insight into the phenomena of interest. Our analysis is limited to statements about collection, retention, usage, and transfer of personal information, which were first studied by Antón and Earp in their seminal paper on privacy goal mining [3].

We prepare the policies for annotation by removing section headers and boilerplate language, and itemizing the policy into individual statements. In each statement, we identify the main data action and categorize the statement into one of five categories: *collection, retention, usage, transfer* and *other*. We only analyze the statements which belong to the first four categories, excluding *others*. Statements that belong to the *others* category are of the following kind, shown with examples from the policy named in parentheses:

- *Definitions* (Costco): "Personal information is information that identifies an individual or that can be reasonably associated with a specific person or entity, such as a name, contact information, Internet (IP) address and information about an individual's purchases and online shopping."
- *User actions* (Barnes and Noble): "You may also access, correct or change the personal information in your community profile(s) on SparkNotes.com at any time, except to change your username."
- *Scope of the privacy policy* (Lowes): "This Privacy Statement applies to the US practices of Lowe's Companies, Inc. and its US operating subsidiaries and affiliates except as outlined below."
- *Customer relations* (Overstock): "If you have questions about your order, you should direct them to us and not to the Vendor."

Next, we use the frame-based markup developed by Breaux and Antón to identify semantic roles associated with different data actions [9]. The tool allows us to use first cycle coding [26] and to segment the statement by identifying the phrases that correspond to roles, while accounting for variability in the statement due to logical conjunctions and disjunctions. The markup is then parsed to generate lists of roles based on each action and syntactic cue, which we discuss later. Consider the following example, which annotated statement using the tool and which is from the Lowes privacy policy:

```
   [[This information] may be used {to [provide a
better-tailored shopping experience]}, |and {for
[<market research, | data analytics, | and system
administration> purposes]}.]
```

The guidelines we use to annotate the statements are as follows:

- *Square brackets* are used to denote role fillers that are required to make the statement grammatically correct. For example, in the statement above, the object `[this information]` is required.
- *Curly brackets* are used to denote clauses that can be removed, which typically correspond to optional roles. For example, {to …} and {for…} curly-bracketed clauses in the statement above can be removed and the sentence would still be grammatically correct; however, if the words "to" and "for" are present, then the nested role values within the square brackets would be required for the statement to make grammatical sense. For instance, in the statement above, if we remove the roles in the "to" and "for" patterns, the statement would become: "This information may be used." Each statement is enclosed in a square bracket to demarcate sentence boundaries.
- *Angular brackets* are used when a phrase or clause contains alternative sub-clauses among which at most one sub-clause is needed to produce a grammatically correct sentence. For example, the phrase "and for" above applies to all phrases inside the angular brackets.

After annotation, we code the extracted phrases in curly brackets using open coding [26] to assign semantic role names to these phrases. Example annotation-coded pairs are as follows:

- `[this information]`: object
- `{to [provide a better-tailored shopping experience]}`: data purpose
- `{for [<market research, | data analytics, | and system administration> purposes]}`: data purposes

In this statement, the lexical and syntactic patterns `to [value]` where *value* is "provide a better-tailored shopping experience," and `for [value]` where *value* is "market research, data analytics, and system administration purposes" are used to specify the data purpose role.

In order to identify the variations in semantic role values (RQ2), we begin with the coded roles values produced by applying the above method, and then we use open coding [26] to categorize the role values for the *condition, source* and *target* roles into different categories. Bhatia & Breaux categorized the purpose role values for the same policies in a prior study [7]. We answer research question RQ3, "what are the different lexical and syntactic triggers that indicate semantic role values?" by extracting all lexical and syntactic patterns from the five annotated policies using the frame-based markup tool [9]. Next, we analyze the results to determine how the same pattern, when used with different data actions, indicates different semantic roles and how different patterns lead to the same semantic role.

B. Semantic Roles Content Analysis Results

In this section, we describe the results to answer RQ1-RQ3. The first research question RQ1 concerns the identification of different semantic roles associated with different categories of data actions. We identified a total of 17 unique semantic roles across the five policies and across the four categories of data actions. The most frequent semantic roles are defined as follows, with the question answered in parentheses (see Appendix A for the complete list of semantic roles):

- *Subject*: The entity which acts on the information. (Who is performing the data action?)
- *Object*: The data on which the action is being performed. The values of this role were information types in our study. (What is being acted upon?)
- *Purpose*: The goal or justification for which the action is performed. (Why is the information being acted upon?)
- *Condition*: The states or events under which the data action will be performed on the information. (When will the data action be performed?)
- *Source*: The provider of the information in a collection action. (From whom is the information collected?)
- *Target*: The recipient of the information in the transfer action. (Who is the data being transferred to?)

Table II presents the frequency of semantic role values for each data action category, across all five policies shown in Table I (see Appendix B for policy wise frequency). Note that some actions have multiple instances of the same semantic role attached to them.

TABLE II. FREQUENCY OF SEMANTIC ROLE VALUES ACROSS DATA ACTION CATEGORIES

Semantic Role	Collect	Retain	Use	Transfer
Total Actions	90	19	85	87
action location	0	1	3	1
comparison	0	0	1	0
condition	36	7	10	49
constraint	3	1	3	2
duration	0	1	0	0
exception	0	1	0	3
hypernymy	7	1	0	1
instrument	5	0	0	2
negation	6	1	4	9
object	90	19	85	86
purpose	14	5	69	10
retention location	0	2	0	0
retention property	0	2	0	0
source	30	0	1	0
subject	85	13	73	74
target	2	0	0	55
time of action	2	1	0	2
Total no. of semantic role values	**280**	**55**	**249**	**294**

In Table II, we observe that all of the collection, retention and usage actions have the object role attached, whereas one of the transfer actions is missing the object role in the Costco privacy policy. In our privacy surveys (see Section IV.B), we observe that the participants were the least willing to share their information for transfer actions, and not clearly specifying what information is transferred can further increase the perceived risk.

In our dataset, 94.4% of collection actions have the subject role attached, followed by usage actions which have the subject role attached 85.9%, and transfer actions which have the subject role attached 85.1%. Only 68.4% of retention actions have an attached subject role. The transfer actions have the condition role attached 55.2% of the time, which was followed by collection and retention actions that have the condition role 40% and 36.8%

of the time, respectively. Only 11.8% of the usage actions have the condition role. A large number of usage actions (81.2%) have the purpose role, whereas only a small number of retention (26.3%), transfer (12.6%) and collection (15.6%) actions have the purpose role attached. We further observe that different action words are used to describe data practices belonging to the same data action category. For example, the action words *log, submit, gather,* and *collect* are all used to describe *collection* practices. The action word *log* is often used when the data collection is implicit, or automated, and occurs when the user is browsing or using the website. For example, in the statement, "Like most web sites, our servers log your IP address, the URL from which you accessed our site, your browser type, and the date and time of your purchases and other activities." The action word *submit*, however, is often used when the user submits their information to the website, for example, "When you place an international order, you will submit personal information (e.g. your name, email address, billing address, and shipping address) and other order-related information to JCPenny through and to servers located in the United States." This can include the user's name, address, and payment details, in contrast to logged information that includes IP address and browser type. Thus, different action words depict subtle differences in which objects are associated and expected, despite being within the same broader category. In Section III.B.1 through III.B.3 below, we describe the results from open coding [26] the role values for condition, source and target roles to answer the second research question (RQ2) which concerns the variations in semantic role values. Bhatia and Breaux previously analyzed the role values for purposes in privacy policies, thus we did not include this role in our analysis [7].

1) Categories of Values for Condition Role

We identified 102 instances of the condition role across the five policies. The condition categories are as follows:

- *First party action*: The data action is conditioned on an action performed by the website company itself.
- *Legal*: The data action is performed, if it is required by law.
- *Merger*: The data action is performed, if the company is part of a merger or acquisition.
- *Scope*: The data action performed is limited by practices described in the privacy policy.
- *Third party action*: The data action is performed in response to an action performed by a third party.
- *User action*: The data action is conditioned on an action performed by the user, or a property that the user possesses.

Table III presents the condition role categories with examples and frequency across all five policies.

2) Categories of Values for Source Role

The source role describes the information provider. We identified 31 source role instances across all five policies, which were categorized using open coding as follows:

- *Technology*: The source of collected information is a device or technology.
- *Third party*: The information about the user is collected from a third-party.
- *User*: The information is collected from the user.
- *Vague*: The source of information is present, but unclear.

Table IV presents the source categories with examples and their frequency across the five policies in our dataset.

TABLE III. CONDITION CATEGORIES

Category	Examples	% Freq.
first party action	only if we identify a biometric match to our database of known shoplifters, in the receipt of automatically collected information	12.9%
legal	if we believe we are required to do so by law, or legal process, as we deem appropriate in response to requests by government agencies	5.9%
merger	as part of any merger or sale of company assets or acquisition, if some or all of our business assets are sold or transferred	8.9%
scope	as permitted by this privacy policy	1.0%
third party	if any of these service providers need access to your personal information, when they no longer need it	2.0%
user	if you choose to connect your mobile device to the free in-store Wi-Fi available at Lowe's stores, if you are under 18	61.4%
vague	as necessary	7.9%

TABLE IV. SOURCE CATEGORIES

Category	Example Role Values	% Freq.
technology	your computer and mobile device, third party cookies	22.6%
third party	third party sources, public sources	38.7%
user	you, children under the age of 13	35.5%
vague	various sources	3.2%

The collection of information from technology, or from third parties is generally automated and the user may be unaware that the collection is taking place. In contrast, information collected from the user can be explicit collection, when the user provides their information to the company directly through a website.

3) Categories of Values for Target Role

We identified 57 instances of the target role, which describes the information recipient in a transfer action, and categorized these instances as follows:

- *First party*: The information is transferred to the first party website company.
- *Third party*: The recipient of the information is a third party.
- *Location*: The target is the location where the information is being transferred.
- *Technology*: The information is being transferred to a technology.
- *Vague*: The target of the information is present, but unclear.

Table V presents the target categories, examples, and frequencies across the five policies in our dataset (see Table I).

TABLE V. TARGET CATEGORIES

Category	Example Role Values	% Freq.
first party	JC Penny, us	7.0%
third party	third parties, issuer of the Mastercard	80.7%
location	countries, globally	3.5%
technology	servers, mobile devices	5.3%
vague	others, anyone	3.5%

Lexical and syntactic patterns are used to coordinate role values in a role phrase or clause. Lexical and syntactic patterns describe how keywords attach to different data actions, and as part of syntactically different statements, they specify similar or different semantic role values. To answer RQ3, we identified 49 patterns, with 380 instances across all five policies. Table VI presents the five most frequent patterns, with example consisting of the semantic role name, followed by a colon and an example role phrase from the policy. For each pattern, we also present the pattern frequency across the five policies.

TABLE VI. LEXICAL AND SYNTACTIC PATTERNS

Pattern	Semantic Roles	%Freq.
to [value]	purpose: to provide location-based services, target: to servers, object: to personally identifiable information	28.4%
if [value]	condition: if Barnes and Noble becomes involved in a merger	8.2%
with [value]	condition: with your consent, object: with other information, target: with other companies	7.9%
when [value]	condition: when you interact with JC Penney	7.6%
from [value]	source: from you, action location: from our files	7.6%

We observe that the same lexical and syntactic pattern is used to specify different semantic roles, when attached to different data actions and across different statements. The semantics conveyed by these patterns changes when attached to different data actions and in different contexts. For example, the syntactic pattern with the keyword `to[value]` can be used to introduce different semantic roles in the context of different data actions:

- `to [data purpose]`

"We will *store and use* this information to administer the programs and services in which you choose to participate, and as permitted by this Privacy Policy."

- `to [target]`

"In addition, we *disclose* certain personal information to the issuer of the MasterCard in connection with the administration of the Barnes and Noble MasterCard program."

In addition, different syntactic patterns can be used to introduce the same semantic role. For example, the syntactic pattern `if [value]` and `depending on [value]` can be used to specify the condition role.

- `if [condition]`

"If Barnes and Noble becomes involved in a merger, acquisition, restructuring, reorganization, or any form of … some or all of its assets personal information and your transaction history may be provided to the entities …"

- `depending on [condition]`

"Depending on how you choose to interact with the Barnes and Noble enterprise we may collect personal information …"

In our dataset, we observed that although the patterns `if[value]` and `depending on[value]` both represent the role condition, they cannot be used interchangeably. This is because in our dataset the semantic role values that occur with `if` are specific and the values occurring with `depending on` are comparatively generic set of conditions, which can take one of many possible values.

Table VII presents the keywords for each of the most frequent roles across the five policies.

TABLE VII. KEYWORDS USED TO SPECIFY DIFFERENT SEMANTIC ROLE VALUES

Semantic role	Keywords Used
Object	along, in conjunction with, to, with
Condition	as, as part of, depending on, even if, if, in, when, with, without, unless
Purpose	in an effort to, for, to, that, so that
Target	between, to, with
Source	across, from, through

The pattern `to[value]` occurs 58 times with usage actions, and in 57/58 times, this pattern coincides with the purpose role. When the pattern is attached to transfer actions, it occurs 36 times and 31/36 times it coincides with a target role. Some of the patterns such as `if[value]`, `depending on[value]`, and `when[value]` are only used to specify the condition role.

IV. SEMANTIC ROLES AND PRIVACY RISK

In this section, we describe the study designs and results for measuring the effect of semantic roles on privacy risk.

A. Privacy Risk Study Design

Research question RQ4 asks, "how does the presence or absence of different semantic roles affect the user's perception of privacy risk?" Fischhoff et al. describe risk as the individual's willingness to participate in an activity [15]. To answer RQ4, we modified the empirical framework developed by Bhatia et al., which uses factorial vignette surveys and multilevel modeling to measure the change in perceived risk due to different factor levels [6]. The modifications include introducing factors that correspond to semantic roles, noting that some sentences will include these factors while others will exclude these factors. Multilevel modeling is a statistical regression model with parameters that account for multiple levels in datasets. In addition, the model limits the biased covariance estimates by assigning a random intercept for each subject [16].

In each vignette, we present participants with a scenario that consists of multiple factors, also called independent variables. In addition, the vignette consists of a risk likelihood level, and a risk acceptance scale [6]. The risk likelihood scale developed by Bhatia et al. is based on construal level theory, which shows that a privacy violation affecting *only one person in your family* is considered psychologically closer and more salient than *only one person in your country* [6, 30]. The privacy risk framework measures the privacy risk as the user's willingness to share their data, which is the dependent variable for the factorial vignette surveys, *willingness to share*`($WtS)`, and is estimated from participant ratings on an eight-point, bipolar semantic scale, labeled at each anchor point: *1=Extremely Unwilling, 2=Very Unwilling, 3=Unwilling, 4=Somewhat Unwilling, 5=Somewhat Willing, 6=Willing, 7=Very Willing and 8=Extremely Willing*. In a post-test, participants answer demographic questions, including their gender, age range, education level, ethnicity and household income.

We conducted three studies to measure the effects of the presence or absence of different semantic roles on privacy risk.

The survey participants were recruited from Amazon Mechanical Turk, had completed ≥5000 Human Intelligence Tasks and had an approval rating of 97% or greater. The surveys were published on Survey Gizmo. We recruited 80 participants for each of the three surveys. Participants were allowed to take each survey only once, and the same participant was allowed to take all three surveys. The participants of the first survey were paid $3 and those of the second and third survey were paid $2.

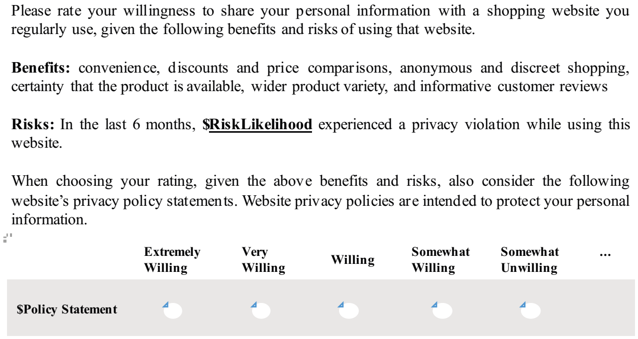
We now describe privacy risk survey modifications.

1) Semantic Roles and Privacy Risk

These studies aim to measure the effect of the presence or absence of different semantic roles across all four data action categories on the perceived privacy risk. To that end, we fixed the values of the *subject* role and *object* role to be "we," and "personal information," respectively. Table VIII presents the factors and corresponding factor level values. Figure 2 presents the factorial vignette survey text.

TABLE VIII. STUDY 1 VIGNETTE FACTORS AND THEIR LEVELS

Factors	Factor Level
Risk Likelihood (`$RL`) Between subject	only one person in your family
	only one person in your workplace
	only one person in your city
	only one person in your state
	only one person in your country
Data actions (`$DA`) Within subject	(C) Collection: collect
	(R) Retention: retain
	(U) Usage: use
	(T) Transfer: share
Semantic Role (`$SR`) Within subject	(DP) Data Purpose: to provide you services
	(Cond.) Condition: when you create an account with us
	(Source) Source: from you



Please rate your willingness to share your personal information with a shopping website you regularly use, given the following benefits and risks of using that website.

Benefits: convenience, discounts and price comparisons, anonymous and discreet shopping, certainty that the product is available, wider product variety, and informative customer reviews

Risks: In the last 6 months, **$RiskLikelihood** experienced a privacy violation while using this website.

When choosing your rating, given the above benefits and risks, also consider the following website's privacy policy statements. Website privacy policies are intended to protect your personal information.

	Extremely Willing	Very Willing	Willing	Somewhat Willing	Somewhat Unwilling	...
$Policy Statement	○	○	○	○	○	

Fig. 2. Template used for vignette generation (fields with $ sign are replaced with values selected from Table VIII and Table IX)

The baseline policy statement for our survey was "We `$DataAction` your personal information," which includes the semantic roles *subject* and *object* associated with the data action. The policy statements `$Policy Statement` for each of the four actions are generated by adding one or more of the semantic roles from Table VIII to the baseline statement. For this survey, we have three different semantic roles, and therefore a total of eight policy statements for each action including the baseline statement, with all combinations of one or more of the semantic roles. For example, the *collection* statement with the roles *data purpose* and *condition* would be: "When you create an account with us, we collect your personal information to provide you services." The second study has the same three dependent variables: risk likelihood, data action and semantic roles. The levels for the *risk likelihood* and *data action* variables are the same for Study 1 and 2. Table IX presents the additional factors and factor levels for the semantic roles used in Study 2.

TABLE IX. STUDY 2 VIGNETTE FACTORS AND THEIR LEVELS

Factors	Factor Level
Semantic Role ($SR) Within subject	(Cond.) Condition: with your consent
	(Source) Source: from you
	(Target) Target for the data action Transfer: third parties

2) Semantic Role Value Categories and Privacy Risk

In the grounded study, we categorized the role values for the *condition*, *source* and *target* roles (see Sections III.A and III.B). The semantic role value categories can affect a user's perception of privacy risk. A user may be more willing to share their information, if the data action is *required by law*, as compared to if the action is performed *as necessary*, which is a vague condition. The most frequent roles in our policy statements after the subject and object roles were condition, source and target. The third study has three pages with all the role value categories for a particular semantic role on each page. Table X presents the factor (a semantic role), the breakout for each semantic role category, followed by the factor levels, which is the semantic role value per category.

TABLE X. STUDY 3 VIGNETTE FACTORS AND THEIR LEVELS

Factors	Category	Factor Level
Condition (`$Cond`) Within subject	first party action	as part of your member profile
	legal action	if we are required to do so by law
	merger action	as part of a merger
	scope	as permitted by this privacy policy
	third party action	if third party service providers need access to your information
	user	with your consent
	vague	as necessary
Source (`$Source`) Within subject	technology	from your computer and mobile device
	third party	from third party sources
	user	from you
	vague	from various sources
Target (`$Target`) Within subject	first party	to us
	third party	to third parties
	location	globally
	technology	to servers
	vague	to others

Re-using the survey design from Figure 2, the `$Policy Statement` is generated by adding the semantic role value category to the baseline statement, "we transfer your personal information" for the *condition* and *target* roles, and "we collect your personal information" for the *source* role.

B. Privacy Risk Survey Results

We now describe our results from three studies described in Section IV.A. above to answer RQ4, which concerns the effect of presence or absence of semantic roles and their values on the user's perception of privacy risk. We report the survey results in two separate series: the first series measures the effect of the four data action categories and the condition, source, purpose and target roles on perceived privacy risk; and the second series measures the changes in privacy risk due to different role values for the condition, source and target roles.

1) Data Action Categories and Semantic Roles

The first and second studies described in Section IV.A.1 measure the effect of the presence and absence of the condition, source, purpose and target roles on the participant's willingness to share their information.

Equation 1 is our main additive regression model for studies 1 and 2 with a random intercept grouped by participant's unique ID (ϵ), the independent within-subjects measure `$RL`, which is the likelihood of a privacy violation, and `$DA`, which is the data action, and `$SR`, which is the semantic role (see Tables VIII and IX). The additive model formula defines the dependent variable `$WtS` (willingness to share) in terms of the intercept α and a series of components, which are the independent variables. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Eq. 1 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience.

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$SR + \epsilon \qquad (1)$$

Tables XI and XII present the results for the baseline statement "We `$DataAction` your personal information." In Tables XI and XII, the row baseline + semantic role(s) presents the value of the coefficient for the statement which is constructed by adding the semantic role(s) to the baseline statement. A positive coefficient signifies an increase in `$WtS` and a negative coefficient represents a decrease in `$WtS` over the baseline.

TABLE XI. Study 1 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Intercept (DataAction-collect)	4.588***	0.378
Risk: only 1 person in your workplace	-0.242	0.524
Risk: only 1 person in your city	-0.697	0.524
Risk: only 1 person in your state	0.197	0.524
Risk: only 1 person in your country	0.021	0.524
Data Action: retain	0.097	0.068
Data Action: transfer	-0.413***	0.068
Data Action: use	0.039	0.068
Baseline+condition	0.006	0.096
Baseline+condition+purpose	0.397***	0.096
Baseline+condition+purpose+source	-0.444***	0.096
Baseline+condition+source	0.016	0.096
Baseline+purpose	0.478***	0.096
Baseline+purpose+source	0.313***	0.096
Baseline+source	-0.794***	0.096

*p≤.05 **p≤.01 ***p≤.001, 4=Somewhat Unwilling

We observe that adding the source role to the baseline statement (e.g., from you) decreases the participant's willingness to share. In addition, specifying the purpose role in any situation increases the willingness to share. Participants were less willing to provide their information when their data can be transferred as compared to when their data is collected by the website. Table XII presents the modeling results for Study 2.

TABLE XII. Study 2 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Intercept (DataAction-collect)	3.795***	0.354
Risk: only 1 person in your workplace	0.078	0.496
Risk: only 1 person in your city	1.340	0.481
Risk: only 1 person in your state	0.791	0.488
Risk: only 1 person in your country	0.088	0.488
Data Action: retain	-0.222	0.088
Data Action: transfer	-1.341	0.088
Data Action: use	-0.328	0.088
Baseline+condition	0.744***	0.088
Baseline+source	0.081	0.088
Baseline+target	-0.141	0.149
Baseline+condition+source	0.784***	0.088
Baseline+condition+target	0.684***	0.149
Baseline+source+target	-0.104	0.149
Baseline+condition+source+target	0.659***	0.149

*p≤.05 **p≤.01 ***p≤.001, 4=Somewhat Unwilling

In Study 2, we observe that adding the condition role, which concerns seeking consent from the user before their data is acted upon, considerably increases the participant's willingness to share their information. In both surveys, we did not observe any statistically significant difference among the levels of the factor *risk likelihood*.

2) Semantic Role Value Categories

We now report results from Study 3 to measure the effect of role values on perceived privacy risk. The policy statements for this survey were generated by adding the role value category to the baseline statement, "we transfer your personal information" for the condition and target roles, and "we collect your personal information" for the source role.

In equations 2.1, 2.2, and 2.3 below we present our main additive regression models for study 3, with a random intercept grouped by participant's unique ID (ϵ), the independent within-subjects measure `$RL`, which is the likelihood of a privacy violation, and `$DA`, which is the data action, and `$Cond` which is the condition role, `$Source` which is the source role, `$Target` which is the target role, (see Table X).

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$Cond + \epsilon \qquad (2.1)$$

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$Source + \epsilon \qquad (2.2)$$

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$Target + \epsilon \qquad (2.3)$$

The baseline for the condition category is "first party," the baseline source is "technology," and the baseline target is "first party." The results appear in Table XIII.

We observe from Table XIII that when information will be transferred on condition of a user consent action, as required by law, or as permitted by the policy, elsewhere, the user's willingness to share increases above the baseline. On the other hand, third-party condition ("if third party service providers need access to your information") decreases the willingness to share below the baseline, whereas the differences between merger and vague condition as compared to the baseline condition are not statistically significant. We observed that the user's willingness to share increases when the information is collected from the user, directly, as compared to when it is collected from their computer or mobile device. With respect to the target role, the user's willingness to share decreases when the information is transferred to third parties, or the target role value is vague.

TABLE XIII. Study 3 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Semantic Role: Condition, baseline: "first party action"		
Intercept (first party)	3.113***	0.355
Condition: legal	1.788***	0.196
Condition: merger	-0.188	0.196
Condition: scope	0.775***	0.196
Condition: third party	-0.875***	0.196
Condition: user	2.213***	0.196
Condition: vague	-0.150	0.196
Semantic Role: Source, baseline: "technology"		
Intercept (technology)	2.325***	0.399
Source: third party	0.100	0.173
Source: user	2.000***	0.173
Source: vague	0.163	0.173
Semantic Role: Target, baseline: "first party"		
Intercept (first party)	3.245***	0.330
Target: location	-1.775***	0.159
Target: technology	-0.050	0.159
Target: third party	-1.438***	0.159
Target: vague	-1.525***	0.159

*p≤.05 **p≤.01 ***p≤.001, 4=Somewhat Unwilling

V. Threats to Validity

Construct validity addresses whether what we measure is actually the construct of interest [33]. To mitigate threats to construct validity, the annotations were performed by one author and then checked by the other author. The privacy risk framework we use for our studies assumes that a person's willingness to share their information corresponds to their acceptance of the risk [6], which was also used in other studies by Acquisti and Knijnenburg to measure risk related to privacy [1, 21]. As noted by Bhatia et al. [6], semantic scale anchor labels used for the dependent variable `$Wts` in the risk surveys could be interpreted differently by participants [10]. To mitigate this threat, we designed our independent factors `$RL`, `$DA`, `$SR`, `$Cond`, `$Source`, and `$Target` as within-subject factors, such that all the participants see and respond to all levels of the independent variables. Subject-to-subject variability is accounted for in our analysis by the random intercept.

Internal validity concerns whether our correlation of the independent and dependent variables is valid [33]. Selecting the number of vignettes to be rated by a participant must take into account multiple factors, including fatigue experienced by the participant, which affects internal validity [31, 27]. We therefore conducted two studies, wherein participants rated different semantic roles and had to rate 32 and 20 statements, respectively, rather than a single study where they had to rate more than 45 statements at one time. In our risk perception studies, we randomized the order of vignettes and the order of questions in each vignette to mitigate confounding effects. We conducted the privacy risk surveys using statements constructed by adding and removing different semantic roles to a baseline statement with the same subject, action, and object. Even though these statements were grammatically correct, they sometimes lacked coherence due to missing contextual information. For example, the statement, "We transfer your personal information, if you are an executive member," is grammatically correct, however, it lacks context to understand executive membership. We limited the context, because additional context can become a confounding factor and affect the risk perception measurements.

The extent to which we can generalize results refers to external validity [33]. We analyzed five privacy policies in this study. We reached saturation in semantic roles after we analyzed the first two privacy policies, Barnes and Noble and Costco. Barnes and Noble policy contained fourteen out of the 17 total semantic roles we identified across all five policies, and Costco contained three additional semantic roles (instrument, retention location, retention property) not present in Barnes and Noble policy. We did not identify any new semantic roles in the other three policies (JCPenny, Lowes, Overstock). Policies not in our dataset and in different domains could contain new semantic roles and syntactic patterns that we did not observe. Similarly, requirements from other domains could contain additional semantic roles. We believe that the list of semantic roles, their categorization and the list of syntactic patterns that we discovered is only complete for our dataset, whereas new policies or requirements documents could require additional analysis. For our risk surveys, our target population is the average U.S. Internet user. As compared to the 2015 PEW Internet and American Life Survey data of US Internet users, the participants that we recruited from Amazon Mechanical Turk had less reported Asian, Black and Hispanic participants [24]. In our risk surveys, 58%-80% of the participants reported their ethnicity as White. Privacy risk perceptions that are affected by ethnicity might therefore be skewed in our study.

VI. Discussion, Conclusion and Future Work

In this paper, we manually annotated and analyzed five privacy policies to identify the different semantic roles and their values attached to the four different categories of data actions: collection, retention, use and transfer. From a total 281 instances of data action, we identified 17 unique semantic roles which occur 878 times. The expected roles for the four categories of data action were subject, information, condition, and purpose. In addition, collection actions frequently have the source role to indicate *from where* the information was collected, and transfer actions have the target role to indicate *to where* the information was transferred. Missing values for these roles in a data practice statement leads to incompleteness in the data practice description and thus become a source of ambiguity. From our analysis, we observe that nearly 32% of retention statements were incomplete with respect to the subject role. In addition, 45% of transfer statements were incomplete with respect to the condition role, and 19% of usage statements were incomplete with respect to the purpose role. We also observed that multiple lexical and syntactic patterns can be used to specify the same semantic role, and in other instances the same pattern can be used to specify different semantic roles. For instance, the pattern `to [value]` specifies a data purpose in 98.3% of instances when attached to a usage action, and specifies a target in 86.1% of instances when attached to a transfer action. Patterns, such as `if [value]` and `when [value]`, are used to specify a condition, irrespective of the action category to which they are attached.

We conducted three studies to measure the effect of semantic roles and role values by category on perceived privacy risk.

From these studies, we observe that describing the purpose for which the user's data will be acted upon considerably increases the user's willingness to share their information. Similarly, specifying that the user's data will be acted upon only under the condition that the user has consented, increases the willingness to share information. In Study 1, adding the source role with the value "from you" decreased the user's willingness to share their information. In this survey, there was no other value of the *source* role. One explanation may be that participants assume that the source suggests the collected information is more sensitive or personal, or that it is collected automatically without user consent. In Study 2, we observed that adding the condition role, which concerns seeking consent from the user before their data is acted upon, considerably increases the participant's willingness to share their information. In Study 2 we also saw an increase in participant's willingness to share their information when the source was added to the baseline statement, as compared to Study 1 where the condition was "when you create an account with us." The participants see multiple statements on the same page in the survey which includes the statements with conditions. The condition value in Study 2 "with your consent" could have primed participants to think about the other statements more positively.

In Study 3, we observe that participant's willingness to share increases when the information is collected from the user directly, as compared to when the information is collected from third parties, or when the source of the information is vague. Participants were also shown multiple sources from which their information could be collected, including from their devices, third parties, and instances where the source role value is vague. These additional sources may have implied that "from you" excludes automated sources in which participants would not be directly involved in the collection process, in other words, there was an anchoring effect. By comparing the sources from which their information is collected, the users may have felt that they have more control over their information, when they directly provide it to the website, as compared to information about them that can be collected by the website from other sources outside their control. Participants were most willing to share their information when they consented to the transfer, or when the transfer was required by law. In addition, participants perceived the least risk when the information was being transferred to the first party company, compared to other targets.

The content analysis technique described in this paper can be used with requirements documents from other domains to build a semantic frame representation for those domains and to consequently identify incompleteness in requirements other than data practices. In future work, we envision using the annotation technique and the findings from this paper to build a corpus of semantic frames for data practices, and then studying ways to develop an automatic role labelling system for privacy policies. The semantic roles that we identified in this study can be used as a starting point to annotate roles in other privacy policies, and to determine when a role dataset reaches saturation. In addition, we believe that the lexical and syntactic patterns that we identified in this paper can be used as features to automate role labeling.

Appendix A: Extracted Semantic Roles

We identified 17 total semantic roles in our analysis, six of which are described in Section III.B. The remaining roles are as follows:

- Action location: The location where the action is performed.
- Comparison: Comparison of the action with other action(s).
- Constraint: The restrictions on the action.
- Duration: The duration for which the action will be performed.
- Exception: Describes an exception to the action.
- Retention property: This role describes how the information is retained. Example role value from Costco policy: *separately from other member databases*.
- Hypernymy: A more generic semantic role value with specific values.
- Instrument: The medium with which the action is performed.
- Negation: The presence of this role signals that the action will not be performed.
- Retention location: The location at which the object of the retention action is retained.
- Time of action: The time at which the action is performed.

Appendix B: Semantic Roles Frequency

The following table presents statistics, including: the total number of data actions identified in each data action category (Total); the number of role value instances for the most frequent roles and the total number of roles attached to each data actions category (Total Roles), for each policy.

TABLE B.I. Frequency of Semantic Roles Across Policies

Policy	Category	Total	Subject	Object	Condition	Purpose	Total Roles
Barnes and Noble	C	30	29	30	16	6	89
	R	7	6	7	4	3	24
	U	22	20	22	4	17	69
	T	24	18	24	12	1	76
Costco	C	16	13	16	4	2	38
	R	4	1	4	0	0	10
	U	16	14	16	5	12	49
	T	28	24	27	20	4	97
JC Penny	C	20	19	20	9	2	69
	R	1	1	1	0	0	2
	U	19	13	19	0	17	51
	T	12	10	12	4	3	40
Lowes	C	14	14	14	3	2	52
	R	5	3	5	2	2	13
	U	12	10	12	0	10	34
	T	15	14	15	10	2	52
Overstock	C	10	10	10	4	2	32
	R	2	2	2	1	0	6
	U	16	16	16	1	13	46
	T	8	8	8	3	0	29
Total		281	245	280	102	98	878

C: Collection, R: Retention, U: Usage, T: Transfer

Acknowledgment

We thank the CMU RE Lab for their helpful feedback. This research was funded by NSF Frontier Award #1330596 and NSF CAREER Award #1453139.

REFERENCES

[1] A. Acquisti, J. Grossklags, "An online survey experiment on ambiguity and privacy," Communications & Strategies, 88(4): 19-39, 2012.

[2] A. Acquisti, S. Gritzalis, C. Lambrinoudakis, and S. di Vimercati, Digital privacy: theory, technologies, and practices, CRC Press, 2007.

[3] A.I. Antón, J.B. Earp, "A requirements taxonomy for reducing web site privacy vulnerabilities," Req'ts Engr. J., 9(3):169-185, 2004.

[4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe, "The Berkeley FrameNet Project," In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (ACL '98), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 86-90.

[5] S. Bellman, E. J. Johnson, S. J. Kobrin, and G. L. Lohse, "International Differences in Information Privacy Concerns: A Global Survey of Consumers," Information Society 20, no. 5 (2004): 313-24.

[6] Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, Thomas B. Norton, "A Theory of Vagueness and Privacy Risk Perception," IEEE 24th International Requirements Engineering Conference (RE'16), Beijing, China, 2016.

[7] J. Bhatia, T.D. Breaux, "A Data Purpose Case Study of Privacy Policies," 25th IEEE International Requirements Engineering Conference, RE: Next! Track, Lisbon, Portugal, 2017.

[8] R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons, "An inference model for semantic entailment in natural language," In National Conference on Artificial Intelligence (AAAI), pages 1678–1679, 2005.

[9] T.D. Breaux and A.I. Antón, "Impalpable constraints: Framing requirements for formal methods," Technical Report Technical Report TR-2006-06, Department of Computer Science, North Carolina State University, Raleigh, North Carolina, February 2007.

[10] L. A. Clark and D. Watson, "Constructing validity: Basic issues in objective scale development," Psychological Assessment, 7(3): 309-319, 1995.

[11] Fabiano Dalpiaz, Ivor van der Schalk, Garm Lucassen, "Pinpointing Ambiguity and Incompleteness in Requirements Engineering via Information Visualization and NLP," Requirements Engineering: Foundation for Software Quality 2018, pp. 119-135.

[12] Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith, "Frame-semantic parsing," Comput. Linguist. 40, 1, March 2014.

[13] Daniel Méndez Fernández, Stefan Wagner, Naming the pain in requirements engineering: A design for a global family of surveys and first results from Germany, Information and Software Technology, Volume 57, 2015, Pages 616-643.

[14] R. E. Fikes and T. Kehler, "The role of frame-based representation in knowledge representation and reasoning," Communications of the ACM 28(9), pp.904-920, 1985.

[15] B. Fischhoff, P. Slovic, S. Lichtenstein, S. Read, B. Combs, "How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits," Policy Sci. 9: 127- 152, 1978.

[16] A. Gelman and J. Hill. 2006, Data analysis using regression and multilevel/hierarchical models, Cambridge Univ. Press, 2006.

[17] J.S. Gruber. Studies in Lexical Relations. Ph.D. thesis, MIT, 1965.

[18] C. J. Fillmore, "Frame Semantics and the Nature of Language," Annals of the New York Academy of Sciences, 280: 20–32.

[19] Daniel Jurafsky and James H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[20] Michael Kaisser and Bonnie Webber, "Question answering based on semantic roles," In Proceedings of the Workshop on Deep Linguistic Processing (DeepLP '07), Association for Computational Linguistics, Stroudsburg, PA, USA, 41-48.

[21] B. Knijnenburg, A. Kobsa, "Increasing sharing tendency without reducing satisfaction: finding the best privacy-settings user interface for social networks," 35th Int'l Conf. Info. Sys., pp. 1-21, 2014.

[22] Gabor Melli, Yang Wang, Yudong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar and Fred Popowich, "Description of SQUASH, the SFU question answering summary handler for the DUC-2005 summarization task," In Proceedings of Document Understanding Conference, Vancouver, Canada, 2005, pp. 103–110.

[23] M. Minsky, "A Framework for Representing Knowledge," J. Haugeland, Ed., Mind Design, MIT Press, 1981.

[24] A. Perrin, M. Duggan, "Americans' Internet Access: 2000-2015," PEW Internet and American Life Project, June 26, 2015.

[25] Michael Roth and Mirella Lapata, "Context-aware Frame-Semantic Role Labeling," Transactions of the Association for Computational Linguistics, v. 3, p. 449-460, August 2015

[26] J. Saldaña, The Coding Manual for Qualitative Researchers, SAGE Publications, 2012.

[27] W. R. Shadish, T.D. Cook, and D. T. Campbell, Experimental and quasi-experimental designs for generalized causal inference, Boston, MA, US: Houghton, Mifflin and Company, 2002.

[28] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth, "Using predicate-argument structures for information extraction," In Proceedings of 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 8-15.

[29] Janice Y. Tsai, Serge Egelman, Lorrie Cranor, Alessandro Acquisti, "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study," Information Systems Research 22(2):254-268.

[30] C. Wakslak and Y. Trope. 2009, "The effect of construal level on subjective probability estimates," Psychol. Sci., vol. 20, no. 1, pp. 52-58, Jan. 2009.

[31] Lisa Wallander, "25 years of factorial surveys in sociology: A review," Social Science Research, Volume 38, Issue 3, 2009, pp. 505-520.

[32] Yinglin Wang, "Semantic information extraction for software requirements using semantic role labeling," 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), Nanjing, 2015, pp. 332-337.

[33] R. K. Yin, Case Study Research: Design and Methods, 5th ed. Sage Pubs., 2013.